THE ASSESSMENT PARADOX

How Testing, Feedback, and Grading Shape Learning

Applied Pedagogy Research Lab

Guido Bartolucci, Principal Investigator

guido@appliedpedagogy.com

LAB.APPLIEDPEDAGOGY.COM

L1-003 · March 2026

Research conducted by AI agents (Claude, Anthropic) under human direction.
See LAB.APPLIEDPEDAGOGY.COM for methodology and verification framework.

CONTENTS

Part I

THE PROBLEM

1

THE PARADOX AT THE HEART OF EDUCATION

Assessment is the most powerful lever available to educators — and the most dangerous one. This is not a metaphor or an overstatement. It is the central empirical finding of thirty years of research: formative assessment practices can produce some of the largest effect sizes in all of education (Black & Wiliam, 1998), the act of testing itself strengthens memory more than additional study time (Karpicke & Roediger, 2008), and well-designed feedback accelerates learning in ways that few other interventions can match (Hattie & Timperley, 2007). At the same time, assessment is the primary mechanism through which schools undermine the intrinsic motivation to learn (Ryan & Weinstein, 2009), the most common form of feedback — grades — is among the least useful and most damaging (Wisniewski, Zierer & Hattie, 2020), and the entire testing apparatus of modern education systematically teaches students that the purpose of learning is to perform on evaluations rather than to develop genuine understanding (Kohn, 1993).

This paradox is not merely theoretical. It plays out every day in every classroom. A teacher who gives a quiz is simultaneously strengthening students' memory (the testing effect), providing information about what they know and don't know (formative assessment), and — depending on how the quiz is framed, graded, and used — either supporting or undermining their desire to learn. The same assessment event can be a powerful learning tool or a motivation killer. The difference lies entirely in the design: the stakes attached, the feedback provided, the framing communicated, and the relationship between assessment and what happens next.

This investigation examines what the evidence actually says about how assessment works, when it helps, when it hurts, and how to design assessment systems that maximize learning without destroying the motivation to keep learning. It covers four major domains — formative assessment, feedback, the testing effect, and alternative assessment — and then confronts the hardest practical question: how do you reconcile the informational value of assessment with its motivational cost?

The investigation builds on the Lo full-field survey, which identified assessment as one of the most powerful practical levers in education but noted that the assessment-motivation intersection remained underresearched (Gap 9). It also builds on the L1-002 motivation investigation, which found robust evidence that extrinsic rewards — including grades, test scores, and rankings — undermine intrinsic motivation, and recommended that assessment practices be redesigned to separate feedback from evaluation wherever possible.

Part II

THE EVIDENCE

2 FORMATIVE ASSESSMENT: THE EVIDENCE AND ITS LIMITS

2.1 THE LANDMARK AND ITS LEGACY

The modern understanding of formative assessment begins with Black and Wiliam's (1998) landmark review, "Assessment and Classroom Learning," published in *Assessment in Education*. The paper has accumulated over 7,400 citations and a field-weighted citation impact of 158.7 — making it one of the most influential papers in the history of education research. Black and Wiliam reviewed approximately 250 studies and concluded that formative assessment — assessment designed to provide feedback that moves learning forward, rather than merely to assign grades — produces substantial learning gains. The effect sizes they reported ranged from 0.4 to 0.7, which, if taken at face value, would make formative assessment one of the most effective educational interventions ever studied.

The core idea is deceptively simple. Formative assessment is any practice through which evidence about student learning is gathered and used — by teachers, by learners, or by peers — to adjust what happens next. It is not a type of test. It is not a product or a technology. It is a process: the process of closing the gap between where learners are and where they need to be. This process requires three things: a clear sense of where the learner is going (learning goals and success criteria), evidence of where the learner currently is (assessment information), and some means of closing the gap (adjusted instruction, feedback, self-regulation).

Black and Wiliam's (1998) paper was not primarily a meta-analysis in the strict statistical sense — it was a narrative review that synthesized findings from a heterogeneous body of studies. This matters because the effect sizes they reported have been questioned by subsequent researchers. Kingston and Nash (2011), in a more methodologically rigorous meta-analysis, found substantially smaller effects — an overall effect size of approximately 0.20, roughly one-third to one-half of what Black and Wiliam reported. The discrepancy is partially explained by methodological differences: Kingston and Nash included only randomized or quasi-experimental studies with comparison groups, excluded several study types that Black and Wiliam had included, and used more conservative aggregation methods.

More recent assessments have landed between these two poles. McMillan, Venable, and Varier (2020) reviewed the state of formative assessment research and concluded that while the evidence base has grown considerably, it remains frustratingly uneven. Studies vary enormously in how they define formative assessment, what outcome measures they use, and what comparison conditions they employ. The most honest summary is this: formative assessment practices, when well-implemented, produce meaningful improvements in learning that are likely in the range of $d = 0.20$ to $d = 0.40$ — smaller than Black and Wiliam's initial estimates but still substantial by educational research standards. The caveat "when well-implemented" is doing significant work in that sentence.

2.2 BLACK AND WILIAM'S FIVE KEY STRATEGIES

In their subsequent theoretical work, Black and Wiliam (2009) refined their framework into five key strategies of formative assessment. These strategies have become the standard operational definition of what formative assessment looks like in practice:

Strategy 1: Clarifying and sharing learning intentions and success criteria. Students cannot assess their own learning or respond effectively to feedback if they do not know what they are aiming for. This strategy involves making learning goals explicit, sharing exemplars of successful work, and helping students develop an understanding of what "quality" looks like in the domain. The theoretical basis is self-regulated learning: learners need a clear target against which to monitor their progress (Nicol & Macfarlane-Dick, 2006).

Strategy 2: Engineering effective classroom discussions, activities, and learning tasks that elicit evidence of student understanding. This is about designing learning activities that generate information about what students know and can do — not just whether they can provide the correct answer, but how they are thinking about the material. Techniques include rich questioning (questions that require explanation rather than recall), concept maps, exit tickets, and think-pair-share activities. The critical feature is that the activities must produce information that can be acted upon, not merely consumed.

Strategy 3: Providing feedback that moves learners forward. This is the most extensively researched of the five strategies and is discussed in detail in Chapter 3. The key principle is that feedback must be actionable — it must tell the learner something specific about what to do differently, not merely evaluate what they have done. A grade is not feedback in this sense. "Good job" is not feedback. "Your argument in paragraph three lacks supporting evidence; here is how you might strengthen it" is feedback.

Strategy 4: Activating students as instructional resources for one another. Peer assessment and peer feedback, when well-structured, can be as effective as teacher feedback for certain purposes — and has the added benefit of developing the assessor's own understanding. When students evaluate another student's work against criteria, they deepen their own understanding of what quality looks like. This connects to Chi's (2009) ICAP framework: interactive learning activities (which include structured peer assessment) produce deeper learning than passive or even individually active ones.

Strategy 5: Activating students as owners of their own learning. This is the self-regulation strategy — developing students' capacity to monitor and direct their own learning. It involves teaching students to self-assess, to generate their own retrieval practice questions, to identify areas of weakness, and to choose appropriate strategies. Andrade's (2019) review of self-assessment research — with a field-weighted citation impact of 120.8 — found that self-assessment can improve learning outcomes, particularly when students are taught how to self-assess effectively using rubrics or criteria.

2.3 WHY FORMATIVE ASSESSMENT IS HARD TO IMPLEMENT

If formative assessment is so effective, why isn't it ubiquitous? The answer lies in a combination of institutional barriers, teacher capacity limitations, and fundamental tensions with the accountability purposes that dominate modern education systems.

Black and Wiliam themselves addressed this in their 2018 paper "Classroom Assessment and Pedagogy," noting that the relationship between assessment and instruction is far more complex than their earlier work had suggested. Formative assessment is not a technique that can be "added on" to existing practice — it requires a fundamental shift in the teacher's role from knowledge

transmitter and evaluator to diagnostician and coach. This shift is cognitively demanding for teachers. It requires them to simultaneously manage instruction, interpret student thinking in real time, and make adaptive decisions about what to do next. Many teachers lack the pedagogical content knowledge to interpret student errors diagnostically — to understand not just that a student got something wrong, but why they got it wrong and what that reveals about their understanding.

The institutional environment often works against formative assessment. When schools are evaluated primarily on summative test scores, the incentive structure pushes teachers toward "teaching to the test" — a form of summative assessment that crowds out the formative practices that would actually improve the learning being tested. This is the testing paradox that Ryan and Weinstein (2009) identified from a self-determination theory perspective: high-stakes accountability systems create controlling environments that undermine both the quality of teaching and the quality of learning.

Wiliam (2011) argued that assessment functions formatively only when the evidence it generates is used to adapt instruction. An assessment can be designed to be formative, but if the teacher collects the results and does nothing with them, or if students receive feedback after it is too late to act on it, the assessment functions summatively regardless of its label. The distinction between formative and summative is not about the instrument — it is about the use to which the results are put. This functional definition is important because it means that any assessment — including standardized tests — can function formatively if the results are used to inform subsequent instruction, and any assessment — including classroom quizzes — can function summatively if the results are used only for grading.

3 FEEDBACK: WHEN IT HELPS, WHEN IT HURTS, AND WHY

3.1 THE SURPRISING FRAGILITY OF FEEDBACK

The common assumption is that feedback improves learning. More feedback is better. Students need to know how they are doing. This assumption is mostly wrong — or at least far too simple.

Kluger and DeNisi (1996) conducted the most comprehensive meta-analysis of feedback interventions to that date, analyzing 607 effect sizes from 131 studies spanning nearly a century of research. Their headline finding was startling: feedback interventions improved performance on average, but decreased performance in more than one-third of cases. Over 38% of the feedback interventions they studied made things worse. This finding should have been more disruptive than it was. It means that the intuitive practice of "giving students feedback" is not reliably beneficial. The effect depends entirely on what kind of feedback, about what, to whom, and in what context.

Kluger and DeNisi proposed Feedback Intervention Theory (FIT) to explain these results. The theory distinguishes three levels at which feedback can direct attention: the task level (how well the specific task was performed), the task-learning process level (how the learner approached the task and what strategies they used), and the self level (what the feedback implies about the learner as a person). The key prediction is that feedback becomes less effective — and potentially harmful — as attention moves from the task level to the self level. When feedback directs attention to the task ("your calculation in step three contains an error"), the learner focuses on the work and how to improve it. When feedback directs attention to the self ("you're really struggling with math"), the learner focuses on their own adequacy, which triggers self-protective processes that interfere with learning.

3.2 THE HATTIE-TIMPERLEY MODEL

Hattie and Timperley (2007) built on Kluger and DeNisi's work to develop the most widely cited framework for understanding feedback in education. Their paper, "The Power of Feedback," has accumulated over 11,500 citations and a field-weighted citation impact of 481.2 — an extraordinary figure that makes it one of the most impactful papers in educational research history.

Hattie and Timperley proposed that effective feedback answers three questions from the learner's perspective:

- **Where am I going?** (Feed up — clarity about goals and success criteria)
- **How am I going?** (Feed back — information about current performance relative to goals)
- **Where to next?** (Feed forward — guidance on what to do to improve)

They further distinguished four levels of feedback, building on Kluger and DeNisi's framework:

Task-level feedback addresses correctness, accuracy, or completeness of the work. It is specific, concrete, and directly actionable. "Your thesis statement in the first paragraph is too broad — it tries to cover three topics. Narrow it to one." This level is highly effective for novice learners working on well-defined tasks where there is a clear standard of correctness. It is less useful for complex, open-ended tasks where there may not be a single correct answer.

Process-level feedback addresses the strategies, approaches, or processes the learner used to accomplish the task. "You jumped straight to calculating without drawing a diagram first. Try representing the problem visually before you start computing." This level is particularly valuable because it develops transferable skills — the learner can apply the strategy to future tasks, not just correct the current one. Process feedback is more effective than task feedback for complex tasks and for more advanced learners.

Self-regulation feedback addresses the learner's capacity to monitor and direct their own learning. "Before you submit your next draft, try reading it aloud to check whether the argument flows logically." This level is most appropriate for learners who have sufficient domain knowledge to engage in meaningful self-monitoring. For novices who lack the schemas to evaluate their own work, self-regulation prompts may be ineffective because the learner does not know what to look for.

Self-level feedback addresses the learner as a person — praise, encouragement, or criticism directed at the self rather than the work. "You're so smart." "You need to try harder." "Great job!" This level is the least effective and potentially the most harmful. Even well-intentioned praise, when it is about the person rather than the work, can shift attention from the task to the self, undermine a mastery orientation, and create dependence on external validation. Kohn (1993) documented this extensively: praise functions as a verbal reward that creates the same motivational dynamics as tangible rewards — temporary compliance coupled with reduced intrinsic interest.

3.3 THE WISNIEWSKI META-ANALYSIS: QUANTIFYING THE DIFFERENCES

Wisniewski, Zierer, and Hattie (2020) conducted an updated meta-analysis of feedback research that provided quantitative support for the Hattie-Timperley framework. Their analysis, published in *Frontiers in Psychology* with a field-weighted citation impact of 199.8, found:

- The overall effect of feedback was $d = 0.48$ — moderate and positive.
- But the variance was enormous. Some feedback conditions produced effects above $d = 1.0$, while others produced negative effects.
- Feedback about the task and about task processing was substantially more effective than feedback about the self.
- Praise and rewards had minimal or negative effects on learning — confirming what Kluger and DeNisi had found two decades earlier.
- The most important moderator was not the timing, frequency, or mode of delivery — it was the content. What the feedback said mattered more than how or when it was delivered.

This finding has a critical practical implication. Much of the conversation about feedback in education focuses on logistics — how quickly to return papers, how often to give quizzes, whether to use written or verbal feedback. These considerations are not irrelevant, but they are secondary to the fundamental question of what the feedback actually communicates. Feedback that directs attention to the task and how to improve it is the most consistently beneficial. Feedback that directs attention to the self — even when that feedback is positive — is the least beneficial and sometimes harmful.

3.4 FEEDBACK TIMING: THE COMPLICATED PICTURE

The question of when to give feedback — immediately or after a delay — turns out to be more complicated than the intuitive answer ("immediately, of course") would suggest.

For simple factual learning and error correction, immediate feedback is generally superior. If a student gives an incorrect answer on a practice quiz, immediate feedback that provides the correct answer prevents the error from being consolidated in memory. The testing effect literature is clear on this point: retrieval practice with immediate corrective feedback produces the best learning outcomes (Roediger & Butler, 2010).

For more complex learning, however, the picture is murkier. There is some evidence that delayed feedback can be more effective for transfer — the ability to apply knowledge in new contexts. The theoretical explanation is that a delay between performance and feedback creates a form of desirable difficulty: the learner must retrieve the original task from memory when processing the feedback, which strengthens the memory trace. Shute's (2008) comprehensive review — with a field-weighted citation impact of 359.0 — concluded that the optimal timing depends on the complexity of the task and the learner's level of knowledge. For simple tasks and novice learners, immediate feedback is best. For complex tasks and more knowledgeable learners, a moderate delay may be more effective.

The practical recommendation is straightforward: provide corrective feedback on factual and procedural tasks as quickly as possible, but for complex tasks where the goal is deep understanding and transfer, a short delay (hours to a day, not weeks) may actually be beneficial — provided the feedback is specific and actionable when it does arrive.

3.5 FEEDBACK SPECIFICITY: THE GOLDILOCKS PROBLEM

How specific should feedback be? The intuitive answer is "as specific as possible." But this too is oversimplified.

Highly specific feedback — telling the learner exactly what is wrong and exactly how to fix it — reduces the cognitive demands on the learner. This can be beneficial for novices who lack the knowledge to diagnose their own errors. But it can be counterproductive for more advanced learners, because it removes the need for the learner to think. If the teacher identifies every error and provides a correction, the student becomes a passive recipient of solutions rather than an active problem-solver. Sadler (2010) argued that conventional feedback practices create a "dangling carrot" problem — students wait for the teacher to tell them what to do rather than developing their own evaluative capacity.

The resolution connects to the expertise reversal effect discussed in the Lo survey. Feedback, like instruction, must be adaptive. Novice learners benefit from specific, directive feedback that tells them what to do. As learners gain competence, feedback should become progressively less directive and more prompting — asking questions, highlighting areas for attention, and encouraging self-diagnosis. The ultimate goal is Nicol and Macfarlane-Dick's (2006) vision: feedback that develops the learner's own capacity for self-assessment, not feedback that creates dependence on external evaluation.

3.6 FEEDBACK LITERACY: THE MISSING COMPETENCE

Carless and Boud (2018) introduced the concept of "feedback literacy" — the capacity to make productive use of feedback. Their paper has accumulated over 1,800 citations and a field-weighted citation impact of 363.4, reflecting the field's growing recognition that giving feedback is only half the challenge. Students must also be able to receive, interpret, and act on feedback — and many cannot.

Feedback literacy involves several competencies: *appreciating feedback* — understanding that feedback is about improving learning, not about personal evaluation; *making judgments* — the ability to evaluate the quality of one's own work against standards; *managing affect* — the emotional dimension of receiving feedback, since negative feedback can trigger defensiveness, anxiety, or disengagement; and *taking action* — converting feedback information into specific changes in approach or behavior.

The feedback literacy framework has important implications for curriculum design. Simply providing better feedback is insufficient if students lack the capacity to use it. Developing feedback literacy requires explicit instruction, modeling, and practice — the same approach that works for developing self-regulation more broadly (as the L1-002 investigation found).

3.7 PEER FEEDBACK: AN UNDERUSED RESOURCE

Peer assessment — having students provide feedback to each other — is one of the most promising and underused assessment practices. When well-structured, peer feedback provides three distinct benefits.

First, it provides additional feedback to the recipient. In a class of 30 students, a single teacher cannot provide detailed, individualized feedback to every student on every assignment. Peer feedback multiplies the feedback capacity of the classroom.

Second, and perhaps more importantly, it develops the assessor's understanding. When students evaluate another student's work against criteria, they must deeply engage with the criteria themselves. This evaluative practice develops the kind of "guild knowledge" that Sadler (2010) argued is essential for academic development — the tacit understanding of what quality looks like in a domain.

Third, peer feedback develops the feedback literacy that Carless and Boud (2018) described. Students who regularly give and receive peer feedback learn to separate feedback from personal judgment, to evaluate work against criteria, and to provide actionable suggestions — all skills that serve their own self-regulation.

The evidence base for peer feedback is positive but conditional. Peer feedback is most effective when students are trained in how to provide feedback, when clear criteria or rubrics guide the process, when the feedback focuses on specific aspects of the work rather than global evaluation, and when there is accountability for the quality of feedback provided (not just the fact that it was given). Without these conditions, peer feedback can devolve into superficial praise ("looks good!") or unhelpful criticism.

4 THE TESTING EFFECT: FROM LABORATORY TO CLASSROOM

4.1 THE LABORATORY FOUNDATION

The testing effect — the finding that practicing retrieval strengthens memory more than additional study time — is one of the most robust findings in cognitive psychology. Roediger and Karpicke (2006) provided the foundational demonstration in a paper published in *Psychological Science*: students who studied a passage and then took a practice test on it retained significantly more information after one week than students who studied the passage twice in the same amount of time. Crucially, students predicted the opposite — they expected that restudying would produce better outcomes. The testing effect is not just real; it is counterintuitive.

Karpicke and Roediger (2008) strengthened the evidence in a *Science* paper — 1,746 citations, FWCI of 24.5 — demonstrating that repeated testing with feedback produced dramatically better long-term retention than repeated studying, even when students devoted the same total time to learning. The effect held across diverse materials, including vocabulary, prose passages, and scientific texts. Karpicke and Blunt (2011), also in *Science*, showed that retrieval practice outperformed elaborative studying with concept mapping — a finding that surprised many educators who had assumed that "deep" processing strategies like concept mapping would be superior.

4.2 META-ANALYTIC EVIDENCE

Three major meta-analyses have quantified the testing effect:

Rowland (2014) analyzed 159 experimental comparisons and found an overall effect of $g = 0.50$ for testing versus restudy conditions — a moderate-to-large effect. The effect was larger when tests included feedback, when the initial test was more difficult (requiring more effortful retrieval), and when the final test used the same format as the practice test (though some transfer to different formats was observed). The effect was robust across different types of material and learner populations.

Adesope, Trevisan, and Sundararajan (2017), in their meta-analysis published in the *Review of Educational Research*, found an overall effect of $g = 0.60$ for practice testing across 272 independent comparisons. They found that the effect was present for factual, conceptual, and application-level outcomes, though it was strongest for factual recall. Multiple-choice practice tests were effective but produced smaller effects than short-answer or free-recall tests — likely because short-answer and free-recall require more effortful retrieval, which produces greater memory strengthening.

Yang, Luo, Vadillo, Yu, and Shanks (2021), in a systematic review and meta-analysis published in *Psychological Bulletin*, focused specifically on the translation of the testing effect from laboratory to classroom settings. This is the key review for the practical question of whether the testing effect "works in the real world." They found an overall effect of $g = 0.49$ in classroom studies — slightly smaller than laboratory effects but still substantial and practically meaningful. The effect held across subject areas (STEM, social sciences, humanities), educational levels (elementary through university), and assessment formats. This finding is particularly important because it addresses the ecological validity concern that plagues much cognitive psychology research: the testing effect is not just a laboratory curiosity — it replicates in real classrooms with real students.

4.3 BOUNDARY CONDITIONS AND NUANCES

The testing effect is robust, but it is not unlimited. Several important boundary conditions have emerged from the research:

Material type matters. The testing effect is strongest for factual and conceptual material — the kind of learning that involves building and retrieving declarative knowledge structures. Its benefits for procedural skills (like solving physics problems or writing essays) are less consistent. This does not mean retrieval practice is useless for procedural learning — it means the testing must be adapted. Practicing retrieval of problem-solving strategies, for example, can enhance procedural learning, but simply recalling facts about how to solve problems is less effective than actually practicing the procedures.

Test format matters. Free-recall tests (where the learner must generate the answer from memory) produce larger testing effects than recognition tests (like multiple-choice, where the answer is provided among options). The explanation is effort: more effortful retrieval produces greater memory strengthening. However, multiple-choice tests with plausible distractors can also produce substantial effects, particularly when accompanied by feedback that explains why incorrect options are wrong.

Feedback is essential for error correction. Retrieval practice without feedback can sometimes strengthen incorrect responses — if a student retrieves the wrong answer, the act of retrieval can consolidate the error. Feedback after retrieval practice serves a dual purpose: it corrects errors and provides an additional study opportunity for material that was not successfully retrieved. The combination of retrieval attempt plus corrective feedback is more powerful than either alone (Roediger & Butler, 2010).

Spacing interacts with testing. The testing effect is enhanced when retrieval practice is distributed over time rather than massed into a single session. Spaced retrieval practice — practicing recall at increasing intervals — is among the most powerful combinations in the learning science toolkit. It leverages both the testing effect (retrieval strengthens memory) and the spacing effect (distributed practice produces more durable learning than massed practice).

Initial success rate matters. If retrieval practice is so difficult that students cannot recall anything, the testing effect disappears — there is nothing to strengthen. Conversely, if retrieval is trivially easy, there is little strengthening because the retrieval is effortless. The optimal initial success rate appears to be in the range of 50–80% — difficult enough to require effort, but not so difficult that the learner fails completely. This connects to the broader concept of desirable difficulties (Bjork, 1994): learning activities that feel harder in the moment but produce better long-term outcomes.

4.4 THE RECONCEPTUALIZATION: TESTS AS LEARNING TOOLS

The most radical implication of the testing effect research is not that tests help students learn — it is that this finding requires a fundamental reconceptualization of what testing is for. In conventional education, tests serve an evaluative function: they measure what students have learned. The testing effect research shows that tests simultaneously serve a learning function: the act of retrieval during a test strengthens the very knowledge being tested.

This means that every hour spent on high-stakes summative testing — testing whose primary purpose is to generate grades or accountability data — is an hour that could instead be spent on low-stakes retrieval practice that actually strengthens learning. The institutional incentive structure has it backwards: the form of testing that serves the institution (high-stakes, infrequent,

summative) is the least beneficial for learning, while the form that serves the learner (low-stakes, frequent, formative) is the most beneficial.

The reconceptualization also reframes the emotional relationship between students and tests. In conventional education, tests are sources of anxiety — events to be feared because they carry consequences (grades, rankings, future opportunities). This anxiety is itself a form of extraneous cognitive load that impairs the very retrieval the test is meant to exercise. Low-stakes retrieval practice, by contrast, removes the anxiety. When students know that the quiz "doesn't count" toward their grade, they can engage in the effortful retrieval that strengthens memory without the cognitive and emotional interference of performance pressure.

This framing connects directly to self-determination theory, as the L1-002 investigation documented. High-stakes testing is experienced as controlling — it tells students what to learn, when to learn it, and punishes failure. This controlling context undermines autonomy and shifts motivation from intrinsic ("I want to understand this") to extrinsic ("I need to pass this test"). Low-stakes retrieval practice, by contrast, can be experienced as informational — it tells students where they are in their learning and helps them identify what they need to work on next. This informational context supports competence (by providing evidence of progress) and can support autonomy (by giving students information they can use to direct their own learning).

Part III

THE TENSION

5 THE ASSESSMENT-MOTIVATION TENSION

5.1 THE PROBLEM

This chapter addresses the defining challenge of this investigation: the tension between assessment's informational value and its motivational cost. Assessment informs learning, but the most common forms of assessment — grades, test scores, rankings — are also the most pervasive system of extrinsic rewards and punishments in education. The L1-002 investigation established that extrinsic rewards undermine intrinsic motivation through a robust and well-replicated mechanism (Deci, Koestner & Ryan, 1999). Grades are extrinsic rewards. Therefore, the standard grading system is, by the logic of the evidence, a mechanism for undermining intrinsic motivation.

This is not a theoretical abstraction. Butler (1988) conducted a classic study that demonstrated the point with unusual clarity. She assigned students to three feedback conditions: comments only (specific task-focused feedback without grades), grades only (a letter grade with no comments), and comments plus grades. Students who received comments only showed the highest levels of subsequent interest and performance. Students who received grades only showed the lowest. And students who received comments plus grades performed at the same level as students who received grades only — the presence of the grade effectively negated the benefit of the comments. When students see a grade, they stop reading the feedback. The grade becomes the message.

This finding has been replicated and extended. Ryan and Weinstein (2009), writing from a self-determination theory perspective, analyzed how high-stakes testing affects the quality of teaching and learning. They found that testing environments characterized by external pressure, controlling contexts, and ego-involving evaluation systematically undermine the basic psychological needs (autonomy, competence, relatedness) that SDT identifies as essential for intrinsic motivation. The more assessment is experienced as controlling — as something done to the student rather than for the student — the more it shifts motivation from intrinsic to extrinsic forms.

Kohn (1993) documented this dynamic extensively in *Punished by Rewards*. His central argument — that the behaviorist logic of "do this and you'll get that" undermines intrinsic motivation, creativity, and genuine learning — is supported by the SDT evidence base. Grades, stickers, honor rolls, class rankings — all function as contingent rewards that create the motivational dynamics the undermining effect predicts: short-term compliance coupled with reduced intrinsic interest and increased performance orientation (focusing on looking good rather than learning deeply).

5.2 THE MECHANISM: WHY GRADES UNDERMINE LEARNING

The mechanism through which grades undermine learning operates through several pathways:

Shifted locus of causality. When students work for grades, the perceived reason for engagement shifts from internal ("I find this interesting" or "I want to understand this") to external ("I need to get a good grade"). This shift — which SDT calls a shift from internal to external perceived locus of causality — reduces the quality of engagement. Students who study for grades are more likely to use surface strategies (memorization, cramming) rather than deep strategies (elaboration, self-explanation) because surface strategies are often sufficient to achieve the grade.

Performance orientation. Grades create what achievement goal theorists call a performance orientation — a focus on demonstrating ability (or avoiding demonstrating inability) rather than on developing competence. Students who are performance-oriented avoid challenge, give up more quickly after failure, and are less willing to take intellectual risks. Students who are mastery-oriented — focused on learning and improvement rather than evaluation — show the opposite pattern. Assessment systems that emphasize grades, rankings, and competition between students foster performance orientation. Assessment systems that emphasize feedback, improvement, and mastery of criteria foster mastery orientation.

Anxiety and cognitive interference. For many students, graded assessments generate anxiety that impairs cognitive performance. Test anxiety is a form of extraneous cognitive load: the anxious student is simultaneously trying to retrieve knowledge and manage the emotional threat of potential failure. This dual demand reduces available working memory capacity and impairs the very performance being measured. The irony is substantial: the measurement tool degrades what it measures.

Reduced risk-taking. When assessment carries high stakes, students rationally avoid risk. They choose easier topics for papers, give safe and conventional answers to open-ended questions, and avoid intellectual experiments that might produce interesting results but could also produce poor grades. This is a direct consequence of the incentive structure: when the penalty for failure is a bad grade, the expected payoff of intellectual risk-taking is negative. Assessment systems that punish failure cannot simultaneously encourage the kind of productive struggle that produces deep learning.

5.3 THE PRODUCTIVE FAILURE CONNECTION

Kapur's (2024) productive failure research provides an important perspective on this tension. Kapur demonstrated that having students attempt to solve problems before receiving instruction — and thus fail in their initial attempts — produces deeper learning than providing instruction first. The struggle and failure, when followed by structured instruction that builds on the failed attempts, creates richer and more transferable understanding.

But productive failure requires specific conditions to work: the failure must be low-stakes (not graded), the classroom culture must normalize struggle, students must have autonomy in how they approach the problem, and the failure must be followed by resolution. These conditions are precisely the conditions that high-stakes grading systems undermine. If the failure counts toward a grade, students will avoid the struggle rather than engage with it. If the classroom culture treats failure as evidence of inadequacy rather than a normal part of learning, students will protect their self-image rather than take intellectual risks.

The productive failure framework thus reinforces the core finding from both the testing effect and the motivation literature: assessment that maximizes learning requires low stakes, tolerance of failure, and a focus on process rather than performance.

6 ALTERNATIVE ASSESSMENT APPROACHES

6.1 THE GRADING PROBLEM

The evidence reviewed above creates a practical problem: grades are deeply embedded in educational institutions, expected by students and parents, required for certification and selection, and entrenched in the grammar of schooling (Tyack & Cuban, 1995). Simply abolishing grades is not feasible for most educational contexts. The question is whether alternatives exist that preserve the informational and accountability functions of grading while reducing the motivational damage.

6.2 STANDARDS-BASED GRADING

Standards-based grading (SBG) replaces traditional letter grades and percentage scores with ratings against specific learning standards. Instead of receiving a "B" on a unit, a student might receive "proficient" on Standard 1 (algebraic reasoning), "developing" on Standard 2 (geometric proof), and "advanced" on Standard 3 (statistical analysis). This approach is more informational than traditional grading because it tells the student (and the teacher and the parents) what the student can and cannot do, rather than collapsing all performance into a single number.

The evidence base for SBG is growing but remains thin. Cain, Medina, Romanelli, and Persky (2021) reviewed the deficiencies of traditional grading systems and the potential of alternatives, concluding that traditional grading conflates effort, compliance, and mastery in ways that obscure actual learning. Standards-based approaches separate these dimensions, providing more diagnostic information. However, large-scale outcome studies comparing SBG to traditional grading are scarce. The theoretical case is strong — SBG is more informational, more transparent, and more aligned with mastery learning principles — but the empirical case relies largely on descriptive studies and case reports.

6.3 UNGRADING

Ungrading — the practice of removing grades entirely, or having students assign their own grades based on self-assessment — has gained attention in higher education, particularly since the publication of Blum's (2020) edited volume. The theoretical logic is straightforward: remove the extrinsic reward (grades) and the undermining effect disappears, allowing intrinsic motivation to flourish.

Kjærgaard, Buhl-Wiggers, and Mikkelsen (2023), in a study of gradeless learning in a Danish university, found that removing grades did not harm academic performance and may have improved it for some students. Hall and Meinking (2022) found that ungrading in a university course increased students' sense of autonomy and shifted their focus from performance to learning. Sorensen-Unruh (2024) argued that the ungrading movement lacks a coherent learning theory and risks conflating the removal of grades with the presence of effective assessment.

This criticism is important. The case against grades does not entail that assessment should be eliminated — it entails that the evaluative function (judging the student) should be separated from the informational function (telling the student what they know and what they need to work on). Ungrading that replaces grades with detailed feedback and self-assessment may be more

effective than either traditional grading or the simple absence of grades. Ungrading that removes grades without replacing them with an alternative feedback system may leave students without the information they need to direct their learning.

The practical limitation of ungrading is that it requires institutional contexts where external accountability demands are minimal. In contexts where grades serve gatekeeping functions — college admissions, professional certification, scholarship allocation — simply removing grades is not feasible. The challenge is to design assessment systems that satisfy institutional accountability requirements while minimizing motivational damage.

6.4 PORTFOLIO ASSESSMENT

Portfolio assessment — collecting a body of student work over time and evaluating it holistically — addresses several limitations of traditional testing. Portfolios can capture growth and development (not just performance at a single point in time), can include diverse types of evidence (written work, projects, reflections, peer feedback), and can involve students actively in selecting and evaluating their own work (which develops self-assessment capacity).

The evidence base for portfolio assessment is positive but limited. Portfolio assessment has been widely implemented in writing instruction, teacher education, and the visual arts, with generally favorable results in terms of student engagement and the quality of work produced. However, portfolios face two significant challenges:

Reliability. Portfolio assessment is inherently subjective. Different evaluators may rate the same portfolio differently, and even the same evaluator may rate inconsistently across portfolios. Achieving acceptable inter-rater reliability requires extensive training and calibration, which is expensive and time-consuming.

Scalability. Portfolios are labor-intensive to evaluate. A teacher who must read and respond to 30 portfolio reflections produces far less feedback per hour than a teacher who grades 30 multiple-choice tests. This scalability constraint limits the feasibility of portfolio assessment in large-enrollment courses and standardized assessment contexts.

These limitations are real but may be partially addressed by technology (e-portfolios with structured rubrics and peer review components) and by reconceiving portfolios as learning tools rather than purely evaluative instruments. When students use portfolios to document their learning process — including mistakes, revisions, and reflections on what they learned from failure — the portfolio becomes a powerful metacognitive tool regardless of how it is formally evaluated.

6.5 COMPETENCY-BASED ASSESSMENT

Competency-based education (CBE) decouples assessment from time — students advance when they demonstrate mastery of defined competencies, not when they have completed a specified number of seat hours. Medical education has the most extensive experience with CBE, following Harden's (1999) foundational articulation of outcome-based education.

The theoretical appeal of CBE is that it aligns assessment with learning: students who have not yet mastered a competency receive additional instruction and practice rather than a failing grade. There is no artificial deadline by which mastery must be achieved, and students who master competencies quickly can advance without waiting for slower peers. This individualization connects to the expertise reversal effect: students who have already developed expertise in an area do not benefit from (and may be harmed by) continued instruction at a level below their competence.

The practical challenges are substantial. CBE requires clearly defined competencies, valid and reliable assessments of those competencies, and flexible institutional structures that allow students to progress at different rates. Most educational institutions are built around the Carnegie unit — the assumption that learning is a function of time spent, not competence achieved. Converting a time-based system to a competency-based system requires restructuring scheduling, staffing, and record-keeping in ways that most institutions are not prepared for.

The evidence base for CBE outcomes is limited and concentrated in medical education. Morcke, Dornan, and Eika (2013) found that the evidence is largely descriptive rather than experimental. There are no large-scale randomized trials comparing CBE to traditional time-based education. The theoretical case is compelling — CBE is more aligned with how learning actually works than time-based progression — but the empirical case remains underdeveloped.

Part IV

SYNTHESIS

7 DESIGNING ASSESSMENT THAT SERVES LEARNING

7.1 PRINCIPLES FROM THE EVIDENCE

The evidence reviewed in this investigation converges on a set of principles for assessment design that maximizes learning while managing the motivational tension:

Principle 1: Maximize formative, minimize summative. The proportion of assessment time devoted to formative purposes (generating feedback that moves learning forward) should vastly exceed the proportion devoted to summative purposes (generating grades for accountability). In practice, this means frequent low-stakes assessments — quizzes, exit tickets, peer reviews, self-assessments — and infrequent high-stakes evaluations. The formative assessments should be designed to leverage the testing effect: they should require retrieval, provide immediate feedback, and be spaced over time.

Principle 2: Separate feedback from evaluation. Wherever possible, give feedback and grades at different times, on different occasions, or through different channels. Butler (1988) showed that grades negate the benefit of comments. When students receive a grade alongside feedback, they attend to the grade and ignore the feedback. Separating the two — providing detailed feedback first, with grades arriving later (or not at all) — allows students to engage with the feedback before the evaluative signal overwhelms it.

Principle 3: Make feedback task-focused and actionable. The Hattie-Timperley model and the Wisniewski et al. meta-analysis are clear: feedback about the task and about task-processing strategies is the most effective. Feedback about the self — including both praise and criticism directed at the person — is the least effective and potentially harmful. Every piece of feedback should tell the student something specific about their work and what to do to improve it.

Principle 4: Develop self-assessment capacity. The ultimate goal of formative assessment is to make the teacher's feedback unnecessary — to develop learners who can assess their own work, identify their own gaps, and direct their own improvement. This requires explicit instruction in self-assessment, practice using rubrics and criteria, peer feedback experiences that build evaluative skills, and progressive transfer of assessment responsibility from teacher to student. Andrade's (2019) review confirms that self-assessment can improve learning, but only when students are taught how to do it.

Principle 5: Normalize failure as information. Assessment systems that punish failure discourage the productive struggle that produces deep learning. Low-stakes retrieval practice should be framed as a learning activity, not an evaluation. When students get questions wrong on a practice quiz, the message should be "this shows you what to study next," not "you didn't study enough." Kapur's productive failure framework provides a model: design assessment experiences where initial failure is expected, normal, and followed by instruction that builds on the failure.

Principle 6: Preserve autonomy. Assessment that is experienced as controlling — surveillance, punishment for non-compliance, external pressure — undermines intrinsic motivation (Ryan & Weinstein, 2009). Assessment that is experienced as informational — feedback that helps students understand where they are and where to go — supports it. Practical ways to preserve autonomy include offering choice in assessment topics or formats, providing rationales for why assessments are structured as they are, involving students in developing assessment criteria, and using invitational

language ("this quiz will help you identify what to review") rather than controlling language ("you must pass this quiz to proceed").

Principle 7: Attend to feedback literacy. Providing excellent feedback is necessary but not sufficient. Students must also develop the capacity to receive, interpret, and act on feedback. This means explicitly teaching students what feedback is for, how to read it, how to manage the emotional response to critical feedback, and how to convert feedback into specific actions. Carless and Boud's (2018) feedback literacy framework provides a roadmap.

7.2 A PRACTICAL MODEL

Combining these principles, a well-designed assessment system for a course or curriculum unit might look like this:

Daily: Brief low-stakes retrieval practice (3–5 minutes). This leverages the testing effect and provides real-time formative information to both teacher and students. No grades. Immediate feedback. Spaced across topics to also leverage the spacing and interleaving effects.

Weekly: A more substantial formative assessment — a problem set, a short writing assignment, a peer review exercise. Feedback is detailed, task-focused, and returned within 24–48 hours. Students have an opportunity to revise based on the feedback. No grades, or if grades are required by institutional context, they are based on improvement or completion rather than accuracy.

Monthly or at unit boundaries: A summative assessment that provides evidence of mastery for accountability purposes. This assessment is preceded by ample formative practice on the same types of tasks. The summative assessment is lower-stakes than traditional exams — it contributes to a final evaluation but is not catastrophically high-stakes. Students who demonstrate mastery proceed; students who do not receive additional instruction and another opportunity.

Continuously: Self-assessment and reflection embedded in the learning process. Students regularly evaluate their own work against criteria, identify areas of strength and weakness, and set specific goals for improvement. This develops the self-regulation capacity that the L1-002 investigation identified as the meta-skill enabling all other learning.

This model is not radical. It is, in essence, what Black and Wiliam have been advocating since 1998. What makes it difficult to implement is not the design — it is the institutional context. Assessment systems are not designed by individual teachers; they are constrained by institutional policies, accreditation requirements, grading scales, and community expectations. Changing assessment at scale requires changing institutions, which brings us back to the grammar of schooling problem that the Lo survey identified. The evidence is clear about what works. The challenge is creating conditions under which what works can be implemented.

8 AN INTEGRATED VIEW

8.1 WHY THESE TRADITIONS HAVE DEVELOPED SEPARATELY

The testing effect and formative assessment literatures have developed in surprisingly separate intellectual traditions. The testing effect research emerged from cognitive psychology — specifically from the memory research tradition of Ebbinghaus, Tulving, and Roediger. It treats testing as a cognitive event that strengthens memory traces. Formative assessment research emerged from educational measurement and classroom practice — the tradition of Black, Wiliam, Sadler, and Nicol. It treats assessment as a social and instructional event that generates information for decision-making.

The two traditions ask different questions and use different methods. Testing effect researchers run laboratory experiments comparing test conditions to restudy conditions, measuring retention after a delay. Formative assessment researchers study classroom practices, often using quasi-experimental designs, measuring a range of learning outcomes and sometimes motivational outcomes as well. The testing effect tradition is primarily concerned with memory; the formative assessment tradition is primarily concerned with the relationship between assessment, teaching, and learning.

8.2 THE INTEGRATION

The integration is straightforward once the traditions are set side by side: frequent, low-stakes formative assessment is the practical implementation of the testing effect in educational settings.

When a teacher gives a brief quiz at the beginning of class, that quiz simultaneously: exercises retrieval, strengthening memory (the testing effect); generates information about what students know and don't know (formative assessment); provides an opportunity for immediate corrective feedback (feedback research); and can be experienced as informational rather than controlling, preserving motivation (SDT).

This convergence means that the evidence base for frequent low-stakes assessment is not just the formative assessment literature or the testing effect literature — it is both. The cognitive mechanism (retrieval strengthens memory) and the instructional mechanism (evidence-based adaptation of teaching) and the motivational mechanism (informational rather than controlling assessment) all point in the same direction: toward assessment that is frequent, low-stakes, retrieval-based, feedback-rich, and embedded in instruction rather than separated from it.

Yang et al.'s (2021) finding that the testing effect translates successfully to classroom settings ($d = 0.49$) provides the critical bridge. The testing effect is not just a laboratory finding — it works in real classrooms, with real teachers, and real students. This means that the cognitive science and the classroom research are converging on the same practical recommendation.

9 CLOSING ASSESSMENT

9.1 WHAT REMAINS UNCERTAIN

This investigation has covered substantial ground, but intellectual honesty requires acknowledging what remains uncertain or contested:

The effect size of formative assessment is debated. Black and Wiliam's (1998) estimates of $d = 0.4$–0.7 are likely overestimates. Kingston and Nash's (2011) estimate of $d \approx 0.20$ may be an underestimate, given their conservative inclusion criteria. The true effect likely falls somewhere in between, but the exact magnitude is uncertain. What is not uncertain is the direction: formative assessment, well-implemented, improves learning.

The assessment-motivation interaction is underresearched at scale. We have strong theoretical reasons (SDT) and some empirical evidence (Butler, 1988; Ryan & Weinstein, 2009) to believe that grading undermines motivation. But large-scale experimental studies comparing graded to ungraded learning environments, with long-term follow-up, essentially do not exist.

Alternative assessment at scale is unproven. Portfolios, competency-based assessment, and standards-based grading all have strong theoretical bases. None has a robust empirical base from large-scale implementations in diverse educational contexts.

The optimal balance between formative and summative assessment is unknown. Everyone agrees that more formative assessment is better. Nobody knows exactly how much summative assessment is too much.

Feedback for complex, ill-structured tasks. Most feedback research has been conducted in the context of well-structured tasks where there is a clear standard of correctness. How to provide effective feedback on creative writing, ethical reasoning, entrepreneurial judgment, or artistic practice is much less understood.

9.2 CONFIDENCE LEVELS

9.2.1 *High Confidence — Build On These*

The testing effect is real and translates to classrooms. Retrieval practice improves learning more than restudying. Effect sizes in classroom settings are approximately $d = 0.5$. This finding is among the most replicated in cognitive psychology.

Feedback effectiveness depends on content, not logistics. Task-focused and process-focused feedback improves learning. Self-focused feedback (including praise) is ineffective or harmful. This finding is supported by multiple meta-analyses spanning three decades.

Formative assessment improves learning when well-implemented. The effect size is debated, but the direction is robust across studies, methods, and contexts.

Grades undermine the motivational benefit of feedback. Butler (1988) showed this directly, and the SDT evidence base provides a strong theoretical explanation.

Extrinsic assessment (grades, rankings) shifts motivation from intrinsic to extrinsic. This is a direct application of the robust undermining effect documented in the SDT literature (Deci et al., 1999).

9.2.2 *Medium Confidence — Proceed With Caution*

Self-assessment can improve learning when taught explicitly. Andrade's (2019) review is positive, but the evidence base is smaller and the conditions for effectiveness are more variable.

Peer feedback provides learning benefits for the assessor as well as the recipient. The evidence is promising but conditional on training and structured rubrics.

Standards-based grading is more informational than traditional grading. The theoretical case is strong. The empirical case is thin.

Feedback literacy is a teachable competence. The concept is well-articulated (Carless & Boud, 2018) but the intervention evidence is still developing.

9.2.3 *Low Confidence — Note But Don't Center*

Ungrading produces better outcomes than traditional grading. The emerging evidence is suggestive but based on small samples, self-selected contexts, and short time frames.

Competency-based assessment is superior to time-based progression. The theoretical logic is compelling. The evidence is almost entirely from medical education and is largely descriptive.

Portfolio assessment can replace traditional testing at scale. Individual implementations are promising. Scalability and reliability challenges are unresolved.

9.3 WHAT A CURRICULUM DESIGNER NEEDS TO KNOW

If you are designing a curriculum and want to know what the assessment evidence says, here is the distillation:

Use assessment as a learning tool, not just a measurement tool. The testing effect means that every quiz, every retrieval exercise, every opportunity for students to practice pulling information from memory is strengthening their learning. Design assessment into the fabric of instruction — not as a separate, periodic event.

Give feedback, not grades. When you must give grades (and institutional reality usually demands it), separate them from feedback in time and space. Give detailed, task-focused feedback first. Let students engage with the feedback. Provide grades later, through a different channel. Never let a grade be the primary message a student receives about their learning.

Make assessment low-stakes and frequent. Frequent low-stakes quizzes leverage the testing effect, provide real-time formative information, reduce test anxiety, and create an informational (rather than controlling) assessment environment. Infrequent high-stakes tests do the opposite on every dimension.

Develop self-assessment capacity. The best feedback is the feedback students give themselves — once they have the evaluative capacity to do so. Teach students to assess their own work against criteria. Use peer feedback to develop evaluative skills. Progressively transfer assessment responsibility from teacher to student.

Normalize failure. The productive failure research, the testing effect research, and the motivation research all converge: the most powerful learning often involves struggle and initial failure. Assessment systems that punish failure discourage the very processes that produce deep learning. Make failure safe. Make it informational. Make it expected.

Be honest about the grading problem. The evidence strongly suggests that conventional grading practices are motivationally harmful and informationally impoverished. The evidence does not yet provide a fully worked-out alternative that satisfies all institutional requirements. Standards-based

grading is theoretically superior. Ungrading is theoretically appealing. Neither has been proven at scale. The honest position is: we know grades are a problem, we have promising alternatives, and we need more evidence about those alternatives before we can recommend wholesale adoption.

Remember that assessment is an institutional question, not just a pedagogical one. The most evidence-based assessment practices in the world cannot survive in an institutional environment that demands compliance with high-stakes standardized testing, percentage-based grading scales, and competitive ranking systems. Assessment reform is institutional reform. This connects to the grammar of schooling problem that Applied Pedagogy must grapple with: the evidence is clear, but the implementation is institutional.

BIBLIOGRAPHY

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, 87(3), 659–701.

Andrade, H. (2019). A critical review of research on student self-assessment. *Frontiers in Education*, 4, 87.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about Knowing* (pp. 185–205). MIT Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21, 5–31.

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 545–567.

Blum, S. D. (Ed.). (2020). *Ungrading: Why Rating Students Undermines Learning (and What to Do Instead)*. West Virginia University Press.

Boud, D., & Molloy, E. (2012). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, 38(6), 698–712.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology*, 58(1), 1–14.

Cain, J., Medina, M. S., Romanelli, F., & Persky, A. M. (2021). Deficiencies of traditional grading systems and recommendations for the future. *American Journal of Pharmaceutical Education*, 86(7), 8850.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1(1), 73–105.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627–668.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest*, 14(1), 4–58.

Hall, E., & Meinking, K. A. (2022). Letting go of grades: Creating an environment of autonomy and a focus on learning for high achieving students. *Teaching & Learning Inquiry*, 10, Article 21.

Harden, R. M. (1999). AMEE Guide No. 14: Outcome-based education. *Medical Teacher*, 21(1), 7–14.

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.

Kapur, M. (2024). *Productive Failure: Unlocking Deeper Learning Through the Science of Failing*. Jossey-Bass.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(6018), 772–775.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, 30(4), 28–37.

Kjærgaard, A., Buhl-Wiggers, J., & Mikkelsen, E. N. (2023). Does gradeless learning affect students' academic performance? A study of effects over time. *Studies in Higher Education*, 49(4), 631–645.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284.

Kohn, A. (1993). *Punished by Rewards: The Trouble with Gold Stars, Incentive Plans, A's, Praise, and Other Bribes*. Houghton Mifflin.

McMillan, J. H., Venable, J. C., & Varier, D. (2020). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment, Research, and Evaluation*, 18, Article 2.

Morcke, A. M., Dornan, T., & Eika, B. (2013). Outcome (competency) based education: An exploration of its origins, theoretical basis, and empirical evidence. *Advances in Health Sciences Education*, 18, 851–863.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218.

Roediger, H. L., & Butler, A. C. (2010). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20–27.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432–1463.

Ryan, R. M., & Weinstein, N. (2009). Undermining quality teaching and learning: A self-determination theory perspective on high-stakes testing. *Theory and Research in Education*, 7(2), 224–233.

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35(5), 535–550.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189.

Sorensen-Unruh, C. (2024). The ungrading learning theory we have is not the ungrading learning theory we need. *CBE—Life Sciences Education*, 23(2), es4.

Tyack, D., & Cuban, L. (1995). *Tinkering Toward Utopia: A Century of Public School Reform*. Harvard University Press.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3–14.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, 147(4), 399–435.